2019-04-24

Discovery of Volatile Biomarkers of Parkinson's Disease from Sebum

Trivedi, DK

10.1021/acscentsci.8b00879
ACS Central Science
American Chemical Society

Discovery of volatile biomarkers of Parkinson's disease from sebum

Drupad K Trivedi[1], Eleanor Sinclair[1], Yun Xu[1,5], Depanjan Sarkar[1], Caitlin Walton-Doyle[1], Camilla Liscio[2], Phine Banks[2], Joy Milne[1], Monty Silverdale[3], Tilo Kunath[4], Royston Goodacre[1,5], Perdita Barran*[1]

[1]Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, Princess Street, Manchester, UK, M1 7DN
[2]Anatune, 4 Wellbrook Way, Girton, Cambridge, UK, CB3 0NA
[3]Department of Neurology, Salford Royal Foundation Trust, Manchester Academic Health Science Centre, University of Manchester, UK
[4]Institute for Stem Cell Research, School of Biological Sciences, The University of Edinburgh, Edinburgh UK, EH16 4UU
[5] Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB

Abstract

Parkinson's disease (PD) is a progressive, neurodegenerative disease that presents with significant motor symptoms, for which there is no diagnostic test. We have serendipitously identified a hyperosmic individual, a 'Super Smeller' that can detect PD by odor alone, and our early pilot studies have indicated that the odor was present in the sebum from the skin of PD subjects. Here, we have employed an unbiased approach to investigate the volatile metabolites of sebum samples obtained non-invasively from the upper back of 64 participants in total (21 controls and 43 PD subjects). Our results, validated by an independent cohort, identified a distinct volatiles-associated signature of PD, including altered levels of perillic aldehyde and eicosane, the smell of which was then described as being highly similar to the scent of PD by our 'Super Smeller'.

Background

Physicians in ancient times, including Hippocrates, Galenus, and Avicenna, used odor as a diagnostic tool. Although the olfactory skills of physicians are not routinely used in modern medicine, it is well documented that a number of conditions, predominantly metabolic and infectious diseases, are associated with a unique odor[1], but there is scant evidence for odors as symptoms of neurodegenerative disorders. To the best of our knowledge this is the first study that demonstrates the use of sebum as biofluid to screen for Parkinson's disease. There have been a small number of metabolomics studies of Parkinson's disease using various biofluids such as blood, faeces, saliva, urine, cerebrospinal fluid as well as insect and mouse models of Parkinson's as described in this recent review by Shao and Le[2], there is no mention of a 'PD odor'.

Joy Milne, a Super Smeller whose husband Les was diagnosed with PD in 1986, has demonstrated a unique ability to detect PD by odor[3]. Joy has an extremely sensitive sense of smell, and this enables her to detect and discriminate odors not normally detected by those of an average olfactory ability. Preliminary tests with t-shirts and medical gauze indicated the odor was present in areas of high sebum production, namely the upper back and forehead, and not present in armpits[3]. Sebum is a waxy, lipid-rich biofluid excreted by the sebaceous glands in the skin, over-production of which known as seborrhea, is a known non-motor symptom of PD[4-5]. Parkinson's skin has recently been shown to contain phosphorylated α-synuclein, a molecular hallmark of PD[6-7]. Identification and quantification of the compounds that are associated with this distinctive PD odor could enable rapid, early screening of PD as well as provide insights into molecular changes that occur as the disease progresses and enable stratification of the disease in future.

Volatile organic compounds (VOCs) are often associated with characteristic odors, although some volatiles may also be odorless. The term 'volatilome' describes the entirety of the volatile organic and inorganic compounds that may originate from any organism, or object, which may be analytically characterized. For any given sample under ambient conditions in a confined environment, collecting, identifying, and measuring molecules in its headspace will then define its volatilome. Such measurements can be performed with thermal desorption gas chromatography mass spectrometry (TD-GC-MS), where a sample is placed in a closed vessel. The sample is then heated to encourage the production of volatiles and the headspace is captured for analysis by GC-MS. Investigation of volatile metabolites using mass spectrometry has proven to be extremely useful in clinical studies[8-11] as well as in the analysis of the consistency and provenance of edible items[12-14]. Recently, TD-GC-MS has been used as a volatilome analysis platform for the detection of compounds from bacteria implicated in ventilator associated pneumonia[10], for differentiation between odors due to human and animal decomposition[15], as well as aerosol detection of the fumes from e-cigarettes[16]. This versatility of TD-GC-MS for samples from many sources renders it highly suitable for use in identifying the metabolites that give rise to the distinct scent of PD. We have established a workflow that starts in clinics with the collections of sebum samples from the upper backs of PD patients along with matched control subjects and progresses to the discovery of disease specific volatile metabolites, the odor of which is confirmed by our Super Smeller. (Figure 1, Supplementary information and Table S1A).

In the current study, VOCs from the sample headspace were measured in two cohorts: a 'discovery' cohort (*n=30)* and a 'validation' cohort (*n*=31), to validate discovered biomarkers[17] (for demographics of each cohort see Table S1B and S5). A third cohort consisting of three drug-naïve PD participants was used for mass spectrometry analysis in conjunction with a human Super Smeller *via* an odor port (Figure 1). This proof of principal study provides the first description of the skin volatilome in Parkinson's disease compared to control subjects.

Study Participants

The participants for this study were part of a nationwide recruitment process taking place at 25 different NHS clinics. The participants for this study were selected at random from these sites. The study was performed in three stages. The first two stages (discovery and validation) consisted of 61 samples (a mixture of control, PD participants on medication and drug naïve PD subjects as shown in Table S1B). The first cohort was used for volatilome discovery, and the second cohort was used to validate the significant features discovered in first cohort. A third cohort consisting of three drug naïve PD participants was used for smell analysis from the Super Smeller. Ethical approval for this project (IRAS project ID 191917) was obtained by the NHS Health Research Authority (REC reference: 15/SW/0354). The metadata analysis for these participants is reported in Table S1B. The study design was as outlined in Figure1.

Sample collection

The sampling involved each subject being swabbed on the upper back with a medical gauze. The gauze with sebum sample from participant's upper back was sealed in background-inert plastic bags and transported to the central facility at the University of Manchester, where they were stored at -80 ^{0}C until the date of analysis.

Analytical Method: TD-GC-MS analysis

Description of the technique

A Dynamic Headspace (DHS) GC-MS method was developed for the analysis of gauze swabs which contained sampled participant sebum. DHS is a sample preparation capability for subsequent GC application using the GERSTEL MultiPurpose Sampler (MPS). DHS extracts and concentrates VOCs from liquid or solid samples. The sample is incubated while the headspace is purged with a controlled flow of inert gas through an adsorbent tube. Once extraction and pre-

concentration are completed, the adsorbent tube is automatically desorbed using the GERSTEL Thermal Desorption Unit (TDU). Analytes are then cryo-focused on the GERSTEL Cool Injection System (CIS) Programmed Temperature Vaporizer (PTV) injector before being transferred to the GC for analysis.

In order to correlate the PD molecular signature to the PD smell, the same setup was used in combination with the GERSTEL Olfactory Detection Port (ODP). The ODP allows detection of odorous compounds as they elute from the GC by smell. In fact, the gas flow is split as it leaves the column between the detector of choice (in our case MS) and the ODP to allow simultaneous detection on the two analytical tools. The additional smell profile information can then be acquired as an olfactogram. Voice recognition software and intensity registration allow direct annotation of the chromatogram.

Method details

Gauze swabs were transferred into 20 mL headspace vials and then analyzed by DHS-TD-GC-MS. For the DHS preconcentration step, samples were incubated for 5 min at 60 ^{0}C before proceeding with the trapping step. Trapping was performed by purging 500 mL of the sample headspace at 50 mL.min^{-1} through a Tenax® TA adsorbent tube kept at 40 ^{0}C (GERSTEL, Germany). Nitrogen was used as purge gas. To release the analytes, the adsorbent trap was desorbed in the TDU in splitless mode. The TDU was kept at 30 ^{0}C for 1 min then ramped at 12 ^{0}C.sec^{-1} to 250 ^{0}C and held for 5 min. Desorbed analytes were cryofocused in the CIS injector. The CIS was operated in solvent vent mode, using a vent flow of 80 mL.min^{-1} and applying a split ratio of 10. The initial temperature was kept at 10 ^{0}C for 2 min, then ramped at 12 ^{0}C.s^{-1} to 250 ^{0}C and held for 10 min. The GC analysis was performed on an Agilent GC 7890B coupled to an Agilent MSD 5977B equipped with high efficiency source (HES) operating in EI mode. Separation was achieved on an Agilent HP-5MS Ultra inert 30 m x 0.25 mm x 0.25 µm column. The column flow was kept at 1 mL.min^{-1}. The oven ramp was programmed as following: 40 °C held for 5 min, 10 °C.min^{-1} to 170 °C, 8 °C.min^{-1} to 250 °C, 10 °C.min^{-1} to 260 °C held for 2 min for a total run time of 31 min. The transfer line to the MS was kept at 300 °C. The HES source was kept at 230 °C and the Quadrupole at 150 °C. The MSD was operated in scan mode for mass range between 30 and 800 *m/z*. For the olfactometry approach, the chromatographic flow was split between the mass spectrometer and the GERSTEL Olfactory Detection Port (ODP3) using Agilent Technologies Capillary Flow Technology (three-way splitter plate equipped with make-up gas). The ODP3 transfer line was kept at 100 °C and humidity of the nose cone was maintained constant.

Data pre-processing and deconvolution

TD-GC-MS data were converted to open source mzXML format using ProteoWizard. Each cohort data were deconvolved separately using eRah package for R[18]. Upon deconvolution, in discovery cohort 207 features and in validation cohort 210 features were assigned to detected peaks. The deconvolved analytes were assigned putative identifications by matching fragment spectra with compound spectra present in Golm database, NIST library and Fiehn GCMS library. In discovery cohort 163 features were assigned an identification and in validation cohort 156 features were assigned an identification. The resulting matrices for each cohort consisted of variables and their respective area under the peak for each sample. All data were normalised for age and total ion count to account for confounding variables (Table S1B).

Statistical analyses

The discovery cohort data included a global analysis of all the detected compounds. PLS-DA modelling was carried out using all the measured features. We have not included PCA results because using this unsupervised clustering method, we were unable to see any clustering of data. We attribute this to the complex nature of metabolomics data especially for volatile metabolites. This results in high dimensionality of the data and it is unrealistic to expect that the separation between PD and controls to be the most dominating variance in the data and thus results in poorer display on PCA/MDS plots. Often supervised modelling is required to train the models to find defined differences by overcoming noise.

The data were log-scaled, and Pareto scaled prior to Wilcoxon-Mann-Whitney analysis, PLS-DA and the production of ROC curves. The PLS-DA was created and executed using MATLAB (2018a)[19-20] and the MATLAB functions are freely available from our in-house cluster toolbox hosted at https://github.com/biospec. ROC curves were generated using the R package called pROC[21]. The samples from both cohorts were combined, thus increasing sample size and providing better statistical power while evaluating the performance of this panel of biomarkers (Figure 2C, Figure S1). ROC curves were generated by Monte-Carlo cross validations (MCCV) using balanced sub-sampling. In each of the MCCV, two thirds of the samples were used to evaluate the feature importance. The top two, three, five, seven and nine important features were then used to build classification models, which were validated using the remaining one third of the samples. The process was repeated 500 times to calculate the average performance and

confidence interval of each model. Classification and feature ranking were performed using a PLS-DA algorithm using two latent variables (Figure 2C).

When performing k-nearest neighbours analysis, *k* was chosen to be 5 given the small sample size, the distance parameter used was Euclidean distance which was used as weights such that closer neighbours of a query point have a greater influence than the neighbours further away. During random forest analysis of the same data, 10 decision trees were grown, and the growth control was achieved by not splitting into subsets smaller than five. SVM model was built using LIBSVM, implemented in e1071 package of R[22], with a linear kernel. The cost (C) and regression loss epsilon ($\in$) were determined by performing a grid search and were set at C=10 and $\in$ was set at 0.10.

Results and discussion

Mass spectrometry data were collected, deconvolved and pre-processed as described in the supplementary information. A partial least squares discriminant analysis (PLS-DA) model was built using the discovery cohort data (Figure 2). The classification accuracy of this model was validated by bootstrapping approach (*n*=1000). The variables contributing to classification (*n*=17) were selected using variable importance in projections (VIP) scores where VIP > 1. We note at this stage that one of the 17 metabolites found is 3,4-dihydroxy mandelic acid, a metabolite of norepinephrine in humans. This catechol is also a metabolite of L-dopa, one of the most commonly prescribed medication for Parkinson's. In this study, 3,4-dihydroxy mandelic acid is observed in both drug naïve participants and control participants indicating its presence may originate from endogenous mandelic acid instead of PD drugs. Norepinephrines including 3,4-dihydroxy mandelic acid are key molecules in the anabolism of brain neurotransmitters. Changes in neurons and neurotransmitters is an extremely well-known characterization of PD[23], for instance the decrease of dopamines, a precursor to 3,4-dihydroxy mandelic acid is a known characterization of PD. It could, therefore, be hypothesized that the presence of endogenous 3,4-dihydroxy mandelic acid could be indicative of altered levels of neurotransmitters in PD.

The measured volatilome in the validation cohort data (from a different population than the discovery cohort) was targeted for the presence or absence of these discovered biomarkers. Out of these 17 metabolites, 13 were also found in the validation cohort data, and nine of these had retention times that allowed us to confidently assign them as identical (Table S2). These nine biomarkers found in both cohorts were selected for further analysis and statistical testing. To

evaluate the performance of these biomarkers, we conducted receiver operating characteristic (ROC) analyses with data from both the discovery cohort and the validation cohort (Figure S1). ROC curves and Wilcoxon-Mann-Whitney tests as well as fold-change calculations on individual metabolites shows four out of these nine metabolites had a similar trend in regulation between the discovery and validation cohorts and their performance was also similar as measured by AUC (Table 1, Figure 3). The results from the combined analysis using both cohorts as a single experiment, indicate increased confidence in the data (*p*-values in Table 1, confidence intervals in Figure S1).

We adhered to the Metabolomics Standards Initiative (MSI) guidelines for data analysis and for assignment of identity to features of interest[17] and all identified features are at MSI level two, which means these are putatively annotated compounds (e.g. without chemical reference standards, based upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries). The compounds perillic aldehyde and eicosane are significantly different between PD and control in both the cohorts ($p < 0.05$): perillic aldehyde was observed to be lower in PD samples whereas eicosane was observed at significantly higher levels. Although hippuric acid and octadecanal were not significantly different ($p > 0.05$), the AUC and box plots (Figure 3) between the two cohorts were comparable and showed similar trends of being increased in PD. Previous studies have reported varying abundances of these compounds in other biofluids (Table 2).

Using an odor port attached to the GC-MS instrument, the Super Smeller identified times at which any smell was present and also more importantly the times at which a specific 'musky' smell of PD was detected. Data were presented in the form of an olfactogram, where the presence and relative intensity of each smell were recorded at its corresponding chromatographic retention time. Olfactogram results obtained from the odor port were overlaid on the respective total ion chromatogram from GC-MS (Figure 4A). There was significant overlap between regions that contained up-regulated compounds and regions in which a smell similar or identical to that of PD scent was present. In the chromatographic trace the region between 19 and 21 min is of particular interest (Figure 4B) since the smell associated with the mixture of analytes in that window was described as "very strong" and "musky". This is the same region where three compounds *viz.* hippuric acid, eicosane and octadecanal have been detected in both cohorts and all three were found to be up-regulated in PD subjects.

In order to validate mass spectrometry led biomarkers and to verify the resultant scent, the candidate compounds listed in Table S2 (*n*=17) were purchased and spiked onto gauze swabs (Table S3). An exploratory study with our Super Smeller was performed in which multiple mixtures of compounds (*n*=5) were spiked onto both blank gauze swabs and swabs that contained control sebum. Two final dispensed volumes of the mixtures were tested (40 µL and 100 µL) and all compounds used were at a single concentration (10 µM). In these blinded tests the Super Smeller grouped the samples in order of PD-like odor. She was able to isolate the swabs with a sebum background matrix and described them as more familiar to the PD-like smell than without control sebum. Further tests utilized control sebum as a background matrix for spiking candidate compounds and a range of concentrations was then selected for testing. Mixtures of the candidate compounds (*n=17)* were prepared at a range of concentrations (10 µM, 5 µM, 0.5 µM, 0.05 µM, 0.005 µM) and presented to the Super Smeller in a second blinded test, she was again asked to rank in order of PD-like smell. These results demonstrated she could detect (although not systematically order) the whole range of concentrations offered, and a concentration between 0.05 µM and 0.5 µM gave her the best response. A validation study consisting of three compound mixtures with significance from the MS analysis aimed to distinguish the combination that best gave rise to the most PD-like smell. Three mixture combinations were chosen at a single concentration (0.5 µM); all candidate compounds (*n=17*), all compounds identified in both the discovery and validation cohorts (*n=9*) and the panel of compounds expressed in same direction and differential between PD and control (*n=4*). The mixture of 9 compounds was consistently described as being most akin to the PD-like odor and was slightly overlapped by description and rank with the mixture of 4 compounds. The mixture of 17 compounds was grouped as the same 'smell' as the other two combinations however were described as significantly weaker. We hypothesize this is due to a lower concentration of each compound in the mixture and thus higher interference from background solvent smell. The results from these tests are depicted in Figure S2 whereby the intensity and correlation to the PD-like smell partitions the groups of samples tested. We do not conclude that these chemicals alone constitute the unique smell associated with PD, rather we demonstrate that they contribute to it.

From results obtained from three independent sets of data, from different people with one underlying factor (i.e. PD) separating them, it was clear that several volatile features were found to be significantly different between control and PD participants. There were no significant

differences observed between PD participants on medication and drug naïve PD participants (p > 0.05 for all measured volatiles), indicating that the majority of the analyzed volatilome and by inference sebum, are unlikely to contain drug metabolites associated with PD medication. In addition, applying machine learning approaches such as k-nearest neighbours, random forest and support vector machines (SVM) did not lead to a classification between drug naïve PD participants and PD participants on medication (results in Table S4).

Perillic aldehyde and octadecanal are ordinarily observed as plant metabolites or food additives. It can be hypothesised that with increased and altered sebum secretion such lipid-like hydrophobic metabolites may be better captured or retained on the sebum-rich skin of PD subjects. Skin disorders in Parkinson's have been observed previously and seborrheic dermatitis (SD) in particular has been flagged as a premotor feature of PD[23]. It has been reported by Arsenijevic and coworkers[5] that PD patients who suffer from SD have increased *Malassezia* density on their skin and commensurate higher lipase activity required metabolically by yeast. This increased lipase activity could correlate with the enhanced production of eicosane, perillic aldehyde and octadecanal as highly lipophilic molecules since *Malassezia* requires specific exogenous lipids for growth. Eicosane is reported as being produced by *Streptomyces* as an anti-fungal agent[24] which also supports its increased presence on the skin of PD sufferers. The effects observed in our study could also signal altered microbial activity on the skin of PD patients that may affect the skin microflora causing changes in the production of metabolites such as hippuric acid[25]. These potential explanations for the change in odor in PD patients all suggest a change in skin microflora and skin physiology that is highly specific to PD.

Conclusion

In conclusion, our study highlights the potential of comprehensive analysis of sebum from PD patients and raises the possibility that individuals can be screened non-invasively based on targeted analysis for these volatile biomarkers. We do acknowledge that the current study is limited with smaller sample size, but the power of this study is a *different* validation cohort that consisted of completely different participants. This validation cohort was able to verify the findings and classification model built using data from our discovery cohort. A larger study with extended olfactory data from human smellers as well as canine smellers in addition to headspace analyses is the next step in further characterizing the PD sebum volatilome. This will enable the establishment of a panel of volatile biomarkers associated with PD and will open new avenues for

stratification as well as facilitate earlier detection of PD and further the understanding of disease mechanisms.

Safety statement: No unexpected or unusually high safety hazards were encountered in the course of this work.

Supplementary information includes description of sampling methodology, analytical methods used and has supplementary figures and tables as well as list of all the PIs from recruitment centers that participated in this study.

Acknowledgment We thank Michael J Fox Foundation and Parkinson's UK for funding this study. We also thank our nationwide NIHR recruitment centers for their enthusiasm and rigor during the recruitment process. We are very grateful to all the participants who took part in this study as well as PIs and nurses across all the recruiting centers that helped us with the recruitment of participants.

Author Contributions DKT, ES, CL, JM and PB performed or contributed to the sample preparation, data collection. DKT and YX performed statistical analysis and validation of data. DKT, ES, JM, RG and PB designed the experiments. PB, RG, MS and TK conceived and supervised the study. DKT wrote the manuscript and all authors commented and contributed to the final version of manuscript.

Competing interests Authors declare no competing interests.

Materials and Correspondence All materials and correspondence request should be addressed to Professor Perdita Barran at perdita.barran@manchester.ac.uk

Figure 1: Schematic outline of the workflow described in this study - from sample collection to biomarker discovery. Parkinson's disease patient samples and control participant samples were collected from 25 sites across the UK using gauze swabs to sample the sebum from the top back region from 64 people. Thermal desorption-gas chromatography-mass spectrometry (TD-GC-MS) analysis was performed alongside olfactory analysis results of which were then combined. Statistical analysis was performed on two independent cohorts. Data from discovery cohort consisting of 30 participants were used to create a partial least squares-discriminant analysis (PLS-DA) model and differential features found as a result were then targeted for presence in a separate validation cohort consisting of 31 participants. The significance of these biomarkers was tested using receiver operating characteristic (ROC) analyses and Wilcoxon-Mann-Whitney test. Finally, four features that showed similar statistical significance and expression on both cohorts were selected for biological interpretation of data.

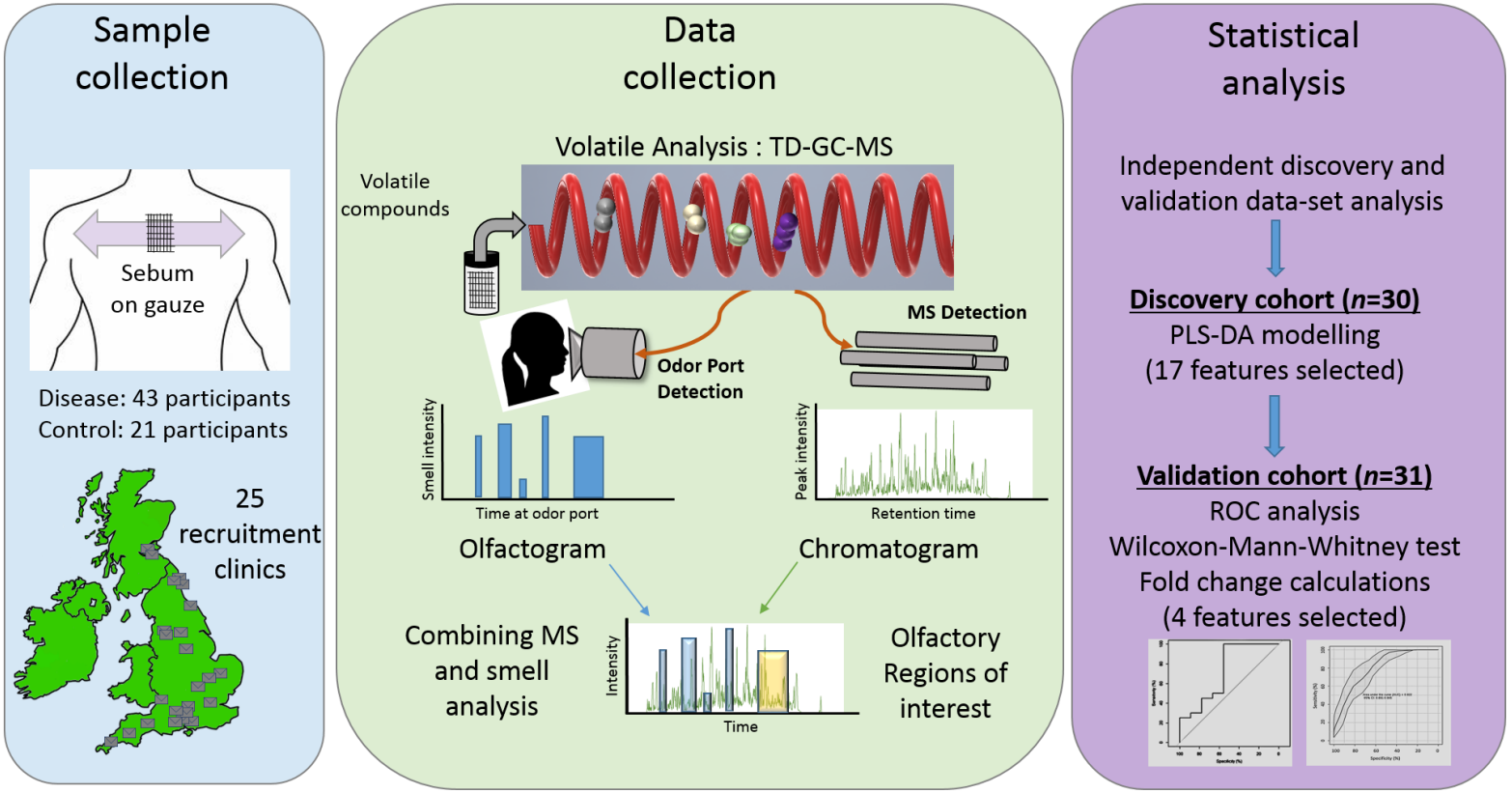
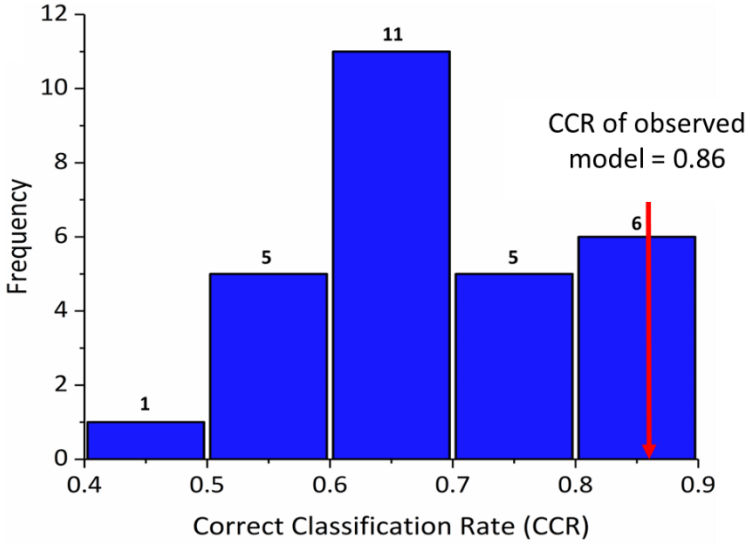


Figure 2: **PLS-DA classification model** (A) Classification matrix of PLS-DA model validated using 5-fold cross validation showing 90% correct prediction of Parkinson's disease samples (B) PLS-DA modelling was further tested using permutation tests (where the output classification was randomised; *n*=26) and results are plotted as a histogram which shows frequency distribution of correct classification rate (CCR) which yielded CCRs ranging between 0.4 to 0.9 for permutated models. The observed model was significantly better than most of the permuted models ($p < 0.1$); shown by the red arrow. (C) ROC plot generated using combined samples from both cohorts and the panel of four metabolites that were common and differential between control and PD. The shaded blue area indicates 95% confidence intervals calculated by Monte Carlo Cross Validation (MCCV) using balanced sub-sampling with multiple repeats.

A

	Observed Control	Observed PD
Predicted Control	0.67	0.33
Predicted PD	0.10	0.90

B



C

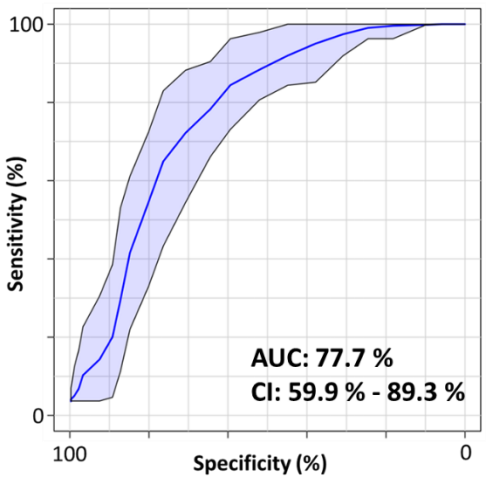


Figure 3: **ROC curves and box plots for analytes of interest:** In each panel from top to bottom: ROC curves for both discovery (blue) and validation (red) cohort for four analytes common to both experiments. Confidence intervals were computed with 2000 stratified bootstrap replicates and diagonal black line represents random guess. Box plots show comparison of means of log scaled peak intensities of these analytes, where black dots were outliers.

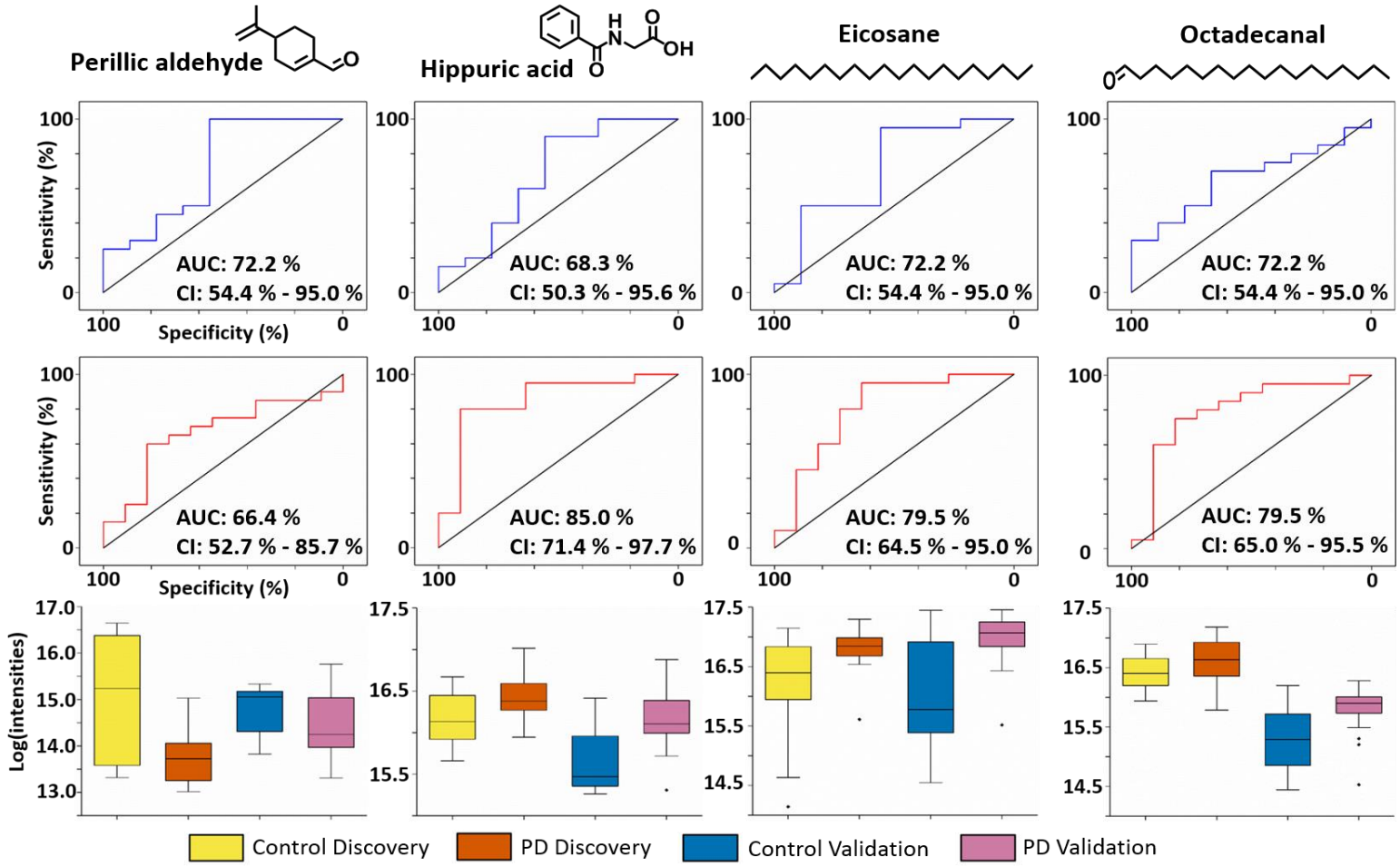


Figure 4: Comparison of GC-MS chromatogram to description of olfactory data described by the Super Smeller: GC-MS chromatogram from three drug naïve Parkinson's subjects and a blank gauze. (A) The 10 to 25 min retention time range of the chromatographic analysis in which the Super Smeller described various odors associated with different GC-MS peaks. The overlaid green shaded area shows the overlap between real time GC-MS analysis and the Super Smeller describing a 'strong PD smell' via the odor port.

(B) A zoom of the green highlighted area from A. This region is of particular interest as 3 out of 4 identified compounds are found here (Tables 1 and S2); it encompasses the time during which the Super Smeller described a musky PD-like scent as being 'very strong' (between the time lines at 19 and 21 min) for the PD samples and not for the blank. It can be noted that none of these compounds are found in blank gauze (bottom chromatogram) within the same retention time window as shown by normalized relative peak intensities to the highest peak in each chromatogram. The area between black dotted lines highlight the presence of compounds in PD samples but complete absence in blank gauze.

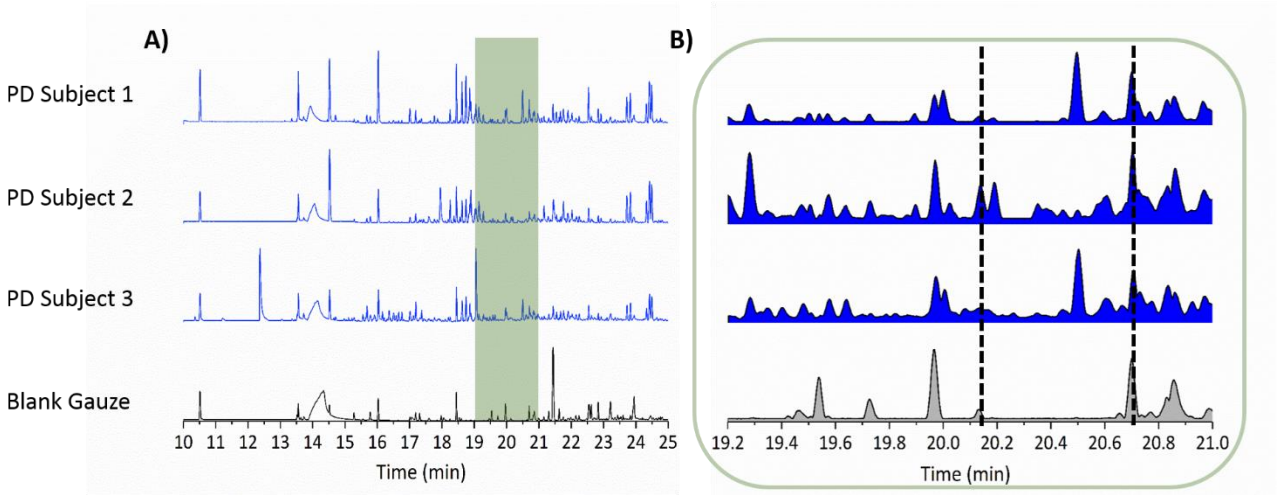


Table 1 : Panel of four volatile metabolites that were found to be differential between Parkinson's and control samples, with similar trends observed in expression and AUC curves measured by

ROC analyses. Perillic aldehyde and Eicosane were significantly down-regulated and up-regulated in PD, respectively (FDR corrected $p < 0.05$) in both cohorts.

Putative identification	Parent Mass	ΔRT (min)	FDR corrected *p*-value (Mann-Whitney test)			Expression (PD/Control)	
			Discovery	Validation	Combined	Discovery	Validation
Perillic aldehyde	150.22	0.15	0.0279	0.0403	<0.0001	Down	Down
Hippuric acid	179.17	0.09	0.1908	0.0403	0.1833	Up	Up
Eicosane	282.56	0.03	0.0279	0.0403	0.0013	Up	Up
Octadecanal	170.34	0.12	0.2605	0.0604	0.3040	Up	Up

Table 2: Known normal abundances of molecules of interest (listed in table 1) measured using mass spectrometry approaches, as reported in literature

Molecule	**Biospecimen**	**Abundance**	**Reference**
Eicosane	Faeces	Not quantified	Garner *et al*[26]
	Saliva	Not quantified	Soini *et al*[27]
	Saliva	Not quantified	Costello *et al*[28]
Hippuric acid	Cerebrospinal fluid	3.0 (0.0-0.5) μM	Hoffman *et al*[29]
	Urine	Not quantified	Hanhineva *et al*[30]
	Blood	16.74 ± 11.16 μM	Duraton *et al*[31]
Octadecanal	Faeces	Not quantified	Dixon *et al*[32]
Perillic aldehyde	Saliva	Not quantified	Costello *et al*[28]

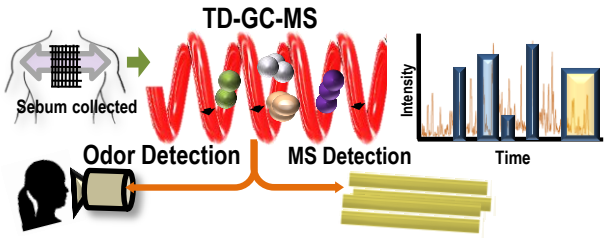
References

1. Shirasu, M.; Touhara, K., The scent of disease: volatile organic compounds of the human body related to disease and disorder. *The Journal of Biochemistry* **2011,** *150* (3), 257-266.
2. Shao, Y.; Le, W., Recent advances and perspectives of metabolomics-based investigations in Parkinson's disease. *Molecular Neurodegeneration* **2019,** *14* (1), 3.
3. Morgan, J., Joy of super smeller: sebum clues for PD diagnostics. *The Lancet Neurology* **2016,** *15*, 138-139.

4. Krestin, D., The Seborrhoeic Facies as a Manifestation of Post-Encephalitic Parkinsonism and Allied Disorders. *QJM: An International Journal of Medicine* **1927,** *os-21* (81), 177-186.
5. Ravn, A.-H.; Thyssen, J. P.; Egeberg, A., Skin disorders in Parkinson's disease: potential biomarkers and risk factors. *Clinical, Cosmetic and Investigational Dermatology* **2017,** *10*, 87-92.
6. Zange, L.; Noack, C.; Hahn, K.; Stenzel, W.; Lipp, A., Phosphorylated alpha-synuclein in skin nerve fibres differentiates Parkinson's disease from multiple system atrophy. *Brain* **2015,** *138* (Pt 8), 2310-21.
7. Donadio, V.; Incensi, A.; Leta, V.; Giannoccaro, M. P.; Scaglione, C.; Martinelli, P.; Capellari, S.; Avoni, P.; Baruzzi, A.; Liguori, R., Skin nerve alpha-synuclein deposits: a biomarker for idiopathic Parkinson disease. *Neurology* **2014,** *82* (15), 1362-9.
8. Bellmann-Weiler, A. P. a. W. F. a. J. W. a. C. A. a. H. W. a. R. K. a. J. B. a. C. T. a. C. A. M. a. G.; Pizzini, A.; Filipiak, W.; Wille, J.; Ager, C.; Wiesenhofer, H.; Kubinec, R.; Blaško, J.; Tschurtschenthaler, C.; Mayhew, C. A.; Weiss, G.; Bellmann-Weiler, R., Analysis of volatile organic compounds in the breath of patients with stable or acute exacerbation of chronic obstructive pulmonary disease. *Journal of Breath Research* **2018,** *12*, 36002.
9. Chang, J.-E.; Lee, D.-S.; Ban, S.-W.; Oh, J.; Jung, M. Y.; Kim, S.-H.; Park, S.; Persaud, K.; Jheon, S., Analysis of volatile organic compounds in exhaled breath for lung cancer diagnosis using a sensor system. *Sensors and Actuators B: Chemical* **2018,** *255*, 800-807.
10. Lawal, O.; Muhamadali, H.; Ahmed, W. M.; White, I. R.; Nijsen, T. M. E. E.; Goodacre, R.; Fowler, S. J.; consortium, o. b. o. t. B., Headspace volatile organic compounds from bacteria implicated in ventilator-associated pneumonia analysed by TD-GC/MS. *Journal of Breath Research* **2018,** *12*, 26002.
11. Rattray, N. J. W.; Hamrang, Z.; Trivedi, D. K.; Goodacre, R.; Fowler, S. J., Taking your breath away: metabolomics breathes life in to personalized medicine. *Trends in Biotechnology* **2014,** *32* (10), 538-548.
12. Xie, J.; Deng, L.; Zhou, Y.; Yao, S.; Zeng, K., Analysis of changes in volatile constituents and expression of genes involved in terpenoid metabolism in oleocellosis peel. *Food chemistry* **2018,** *243*, 269-276.
13. Gatzias, I. S.; Karabagias, I. K.; Kontakos, S. P.; Kontominas, M. G.; Badeka, A. V., Characterization and differentiation of sheep's milk from Greek breeds based on physicochemical parameters, fatty acid composition and volatile profile. *Journal of the science of food and agriculture* **2018,** *68*, 88.
14. Gerhardt, N.; Birkenmeier, M.; Schwolow, S.; Rohn, S.; Weller, P., Volatile-Compound Fingerprinting by Headspace-Gas-Chromatography Ion-Mobility Spectrometry (HS-GC-IMS) as a Benchtop Alternative to 1H NMR Profiling for Assessment of the Authenticity of Honey. *Analytical Chemistry* **2018,** *90*, 1777-1785.
15. Rosier, E.; Cuypers, E.; Dekens, M.; Verplaetse, R.; Develter, W.; Van de Voorde, W.; Maes, D.; Tytgat, J., Development and validation of a new TD-GC/MS method and its applicability in the search for human and animal decomposition products. *Analytical and bioanalytical chemistry* **2014,** *406*, 3611-3619.
16. Herrington, J. S.; Myers, C., Electronic cigarette solutions and resultant aerosol profiles. *Journal of chromatography. A* **2015,** *1418*, 192-199.
17. Goodacre, R.; Broadhurst, D.; Smilde, A. K.; Kristal, B. S.; Baker, J. D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A.; Ebbels, T.; Kell, D. B.; Manetti, C.; Newton, J.; Paternostro, G.; Somorjai, R.; Sjöström, M.; Trygg, J.; Wulfert, F., Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **2007,** *3* (3), 231-241.
18. Domingo-Almenara, X.; Brezmes, J.; Vinaixa, M.; Samino, S.; Ramirez, N.; Ramon-Krauel, M.; Lerin, C.; Díaz, M.; Ibáñez, L.; Correig, X.; Perera-Lluna, A.; Yanes, O., eRah: A Computational

Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics. *Analytical Chemistry* **2016,** *88* (19), 9821-9829.
19. Xu, Y.; Muhamadali, H.; Sayqal, A.; Dixon, N.; Goodacre, R., Partial Least Squares with Structured Output for Modelling the Metabolomics Data Obtained from Complex Experimental Designs: A Study into the Y-Block Coding. *Metabolites* **2016,** *6* (4), 38.
20. Xu, Y.; Goodacre, R., On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing* **2018,** *2* (3), 249-262.
21. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M., pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **2011,** *12* (1), 77.
22. Chang, C.-C.; Lin, C.-J., LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011,** *2* (3), 1-27.
23. Felix-Martin Werner, R. C., Classical Neurotransmitters and Neuropeptides involved in Parkinson's disease: Focus on anti-Parkinsonian Drugs. *Current Drug Therapy* **2015,** *10* (2), 66-81.
24. Ahsan, T.; Chen, J.; Zhao, X.; Irfan, M.; Wu, Y., Extraction and identification of bioactive compounds (eicosane and dibutyl phthalate) produced by Streptomyces strain KX852460 for the biological control of Rhizoctonia solani AG-3 strain KX852461 to control target spot disease in tobacco leaf. *AMB Express* **2017,** *7*, 54.
25. Wikoff, W. R.; Anfora, A. T.; Liu, J.; Schultz, P. G.; Lesley, S. A.; Peters, E. C.; Siuzdak, G., Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences* **2009,** *106* (10), 3698-3703.
26. Garner, C. E.; Smith, S.; Costello, B. d. L.; White, P.; Spencer, R.; Probert, C. S. J.; Ratcliffe, N. M., Volatile organic compounds from feces and their potential for diagnosis of gastrointestinal disease. *The FASEB Journal* **2007,** *21* (8), 1675-1688.
27. Soini, H. A.; Klouckova, I.; Wiesler, D.; Oberzaucher, E.; Grammer, K.; Dixon, S. J.; Xu, Y.; Brereton, R. G.; Penn, D. J.; Novotny, M. V., Analysis of Volatile Organic Compounds in Human Saliva by a Static Sorptive Extraction Method and Gas Chromatography-Mass Spectrometry. *Journal of Chemical Ecology* **2010,** *36* (9), 1035-1042.
28. de Lacy Costello, B.; Amann, A.; Al-Kateb, H.; Flynn, C.; Filipiak, W.; Khalid, T.; Osborne, D.; Ratcliffe, N. M., A review of the volatiles from the healthy human body. *Journal of Breath Research* **2014,** *8* (1), 014001.
29. Hoffmann, G. F.; Meier-Augenstein, W.; Stöckler, S.; Surtees, R.; Rating, D.; Nyhan, W. L., Physiology and pathophysiology of organic acids in cerebrospinal fluid. *Journal of Inherited Metabolic Disease* **1993,** *16* (4), 648-669.
30. Hanhineva, K.; Brunius, C.; Andersson, A.; Marklund, M.; Juvonen, R.; Keski-Rahkonen, P.; Auriola, S.; Landberg, R., Discovery of urinary biomarkers of whole grain rye intake in free-living subjects using nontargeted LC-MS metabolite profiling. *Molecular Nutrition & Food Research* **2015,** *59* (11), 2315-2325.
31. Duranton, F.; Cohen, G.; De Smet, R.; Rodriguez, M.; Jankowski, J.; Vanholder, R.; Argiles, A.; European Uremic Toxin Work, G., Normal and pathologic concentrations of uremic toxins. *Journal of the American Society of Nephrology : JASN* **2012,** *23* (7), 1258-1270.
32. Dixon, E.; Clubb, C.; Pittman, S.; Ammann, L.; Rasheed, Z.; Kazmi, N.; Keshavarzian, A.; Gillevet, P.; Rangwala, H.; Couch, R. D., Solid-phase microextraction and the human fecal VOC metabolome. *PloS one* **2011,** *6* (4), e18471-e18471.

For Table of Contents Use Only



Synopsis:

Sebum from upper-back, analyzed using mass spectrometer hyphenated to an odour port reveals a unique volatilome associated to Parkinson's disease (PD) smell, useful for diagnosing PD non-invasively.

Discovery of volatile biomarkers of Parkinson's disease from sebum

Drupad K Trivedi[1], Eleanor Sinclair[1], Yun Xu[1], Depanjan Sarkar[1], Camilla Liscio[2], Phine Banks[2], Joy Milne[1], Monty Silverdale[3], Tilo Kunath[4], Royston Goodacre[1,5], Perdita Barran*[1]

[1]Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, Princess Street, Manchester, UK, M1 7DN
[2]Anatune, 4 Wellbrook Way, Girton, Cambridge, UK, CB3 0NA
[3]Department of Neurology, Salford Royal Foundation Trust, Manchester Academic Health Science Centre, University of Manchester, UK
[4]Institute for Stem Cell Research, School of Biological Sciences, The University of Edinburgh, Edinburgh UK, EH16 4UU
[5] Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB

Supplementary Information

Table S1A Details of the collecting sites in the UK and the lead PI at each site

SITE NUMBER	SITE NAME	PI
1	Addenbrookes (Cambridge)	Paul Worth
2	Bournemouth	Khaled Amar
3	Cornwall/Truro	Christine Schofield
4	Lothian - Western General Edinburgh	Gordon Duncan
5	Edinburgh – Royal Infirmary of Edinburgh	Gordon Duncan
6	Edinburgh - Primary Care NHS Lothian (Seb Derm)	Richard Weller
7	Hampshire	Sam Arianayagam
8	Nottingham	Gill Sare
9	Pennine	Jason Raw
10	Salford	Monty Silverdale
11	Salisbury	Diran Padiachy
12	Sheffield	Oliver Bandmann
13	South Tees	Neil Archibold
14	Southern Health	Helen Roberts
15	Luton & Dunstable	Anette Schrag
16	Portsmouth	Sean Slaght
17	Northumbria	Richard Walker

18	London North West	Sophie Molloy
19	Bath	Veronica Lyell
20	Gateshead	Richard Athey
21	Sunderland	Uma Nath
22	Plymouth	Camille Caroll
23	Newcastle Upon Tyne Hospitals NHS Foundation Trust (Newcastle University)	Nicola Pavese
24	Royal Devon and Exeter NHS Foundation Trust	Robert James
25	Imperial College Healthcare NHS Trust	Sophie Molloy

Table S1B: Participant numbers and metadata per wave.

		Discovery cohort		
	Control (*n*=10)	Drug Naïve PD (*n*=10)	PD on medication (*n*=10)	*p*-value
Age (years)	64.8 ± 3.06	72.82 ± 8.42	64.67 ± 2.55	0.01*
BMI	27.10 ± 3.50	26.94 ± 4.08	25.33 ± 3.44	0.64
Gender (M/F ratio)	0.84	1.20	0.80	0.88
Alcohol intake (yes/no ratio)	4.5	0.37	2	0.03*
Smoker	1	0	0	0.39
		Validation cohort		
	Control (*n*=11)	Drug Naïve PD (*n*=11)	PD on medication (*n*=9)	*p*-value
Age (years)	55.78 ± 18.87	75.40 ± 6.85	68.90 ± 11.76	0.02*
BMI	28.96 ± 11.01	25.74 ± 3.83	24.98 ± 3.54	1.00
Gender (M/F ratio)	0.26	1.50	1	0.10
Alcohol intake (yes/no ratio)	0.8	9	1.5	0.10
Smoker	0	0	1	0.24

Wave 3 (odor port validation, drug naïve PD subjects only, *n*=3)	
Age (years)	65.66 ± 3.30
BMI	23.46 ± 1.80
Gender (M/F ratio)	2
Alcohol intake (yes/no ratio)	2
Smoker	0

* indicates significant difference between controls, drug naïve and PD with medication groups.

Table S2: List of candidate volatiles putatively identified (MSI level 2) and matched across two different cohorts. Nine of out 17 metabolites listed were selected for further analysis since they had acceptable retention time drift between the two sets of experiments.

Putative identification	Mass	Retention time (min) (discovery)	Retention time (min) (validation)	Retention time (min) difference	Comments
3,4-dihydroxy mandelic acid	184.15	20.87	Not found	n/a	Not found
Artemisinic acid	234.34	12.97	12.83	0.14	Included
Cyclohexasiloxane, dodecamethyl	357.57	16.47	16.06	0.41	Excluded
Cyclohexylcyclohexane	357.57	15.36	14.71	0.65	Excluded
Dodecane	170.34	13.20	13.27	-0.07	Included
Eicosane	282.56	20.65	20.62	0.03	Included
Gallic acid ethyl ester	198.17	11.40	10.99	0.41	Excluded
Glutamine	128.09	21.73	21.09	0.64	Excluded
Hexyl acetate	170.34	11.70	11.53	0.16	Included
Hippuric acid	179.17	20.61	20.52	0.09	Included
Neoabietic acid	302.46	21.66	Not found	n/a	Not found
Octacosane	394.77	17.49	17.46	0.03	Included
Octadecanal	170.34	20.87	20.75	0.12	Included
Octanal	244.38	11.58	11.32	0.26	Included
Perillic aldehyde	150.22	11.82	11.66	0.15	Included
Proline	115.13	14.27	13.77	0.50	Excluded
Tetracosane	338.65	18.17	Not found	n/a	Not found

Table S3: Various standards used to create chemical mixtures that were spiked on gauze containing human sebum, were made by dissolving them in appropriate solvents. The table shows solvents in which each of these standards were individually created, before forming mixtures at various concentrations used for validation of smell by the Super Smeller.

Standard	Solvent
3,4-dihydroxy mandelic acid	Water
Cyclohexasiloxane, dodecamethyl	Methanol
Cyclohexylcyclohexane	Methanol
Diglycerol	Water
Dodecane	Ethanol
Eicosane	Acetone
Gallic acid ethyl ester	Water
Glutamine	Water
Hexyl acetate	MeOH
Hippuric Acid	Water
Hydroxymyristic acid	Dichloromethane
Octacosane	Chloroform
Octanal	Methanol
Perillic aldehyde	Water
Proline	Water

Table S4: Classification approaches attempted to distinguish between drug naïve PD participants and PD participants on medication did not show a very clear classification between the two groups in discovery cohort and validation cohort. *K*-nearest neighbours, random forest and support vector machines (SVM) classification algorithms were used. Using random sampling repeated 10 times, 60% data were used for training set and remaining data were used to test the model. Area under the curve (AUC) and averaged classification accuracy (ACA) for each model are shown.

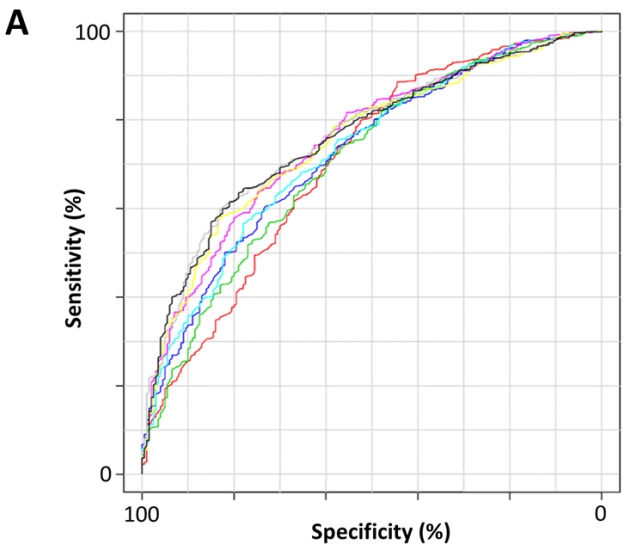
	Discovery cohort		Validation cohort	
Method	**AUC**	**ACA**	**AUC**	**ACA**

kNN	78%	61%	57%	42%
SVM	65%	60%	60%	35%
Random Forest	66%	61%	54%	38%

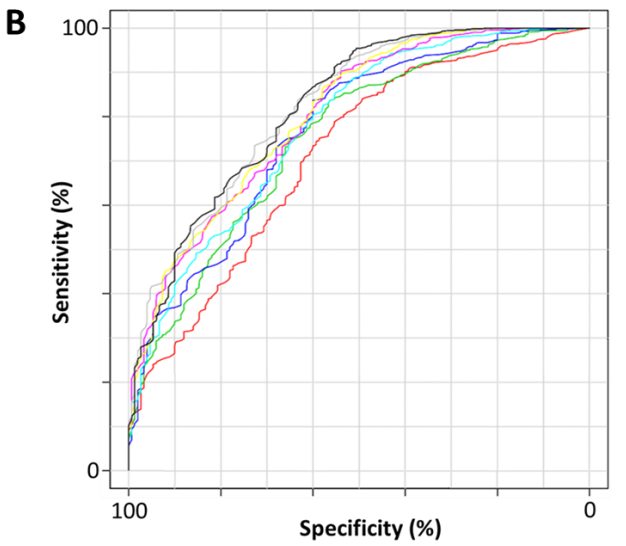
Table S5: List of participants' anonymized ID along with time since they were diagnosed with Parkinson's as of the date of recruitment to this study. Median time since diagnosis for those in Drug Naïve group was 0 year whereas those in Medication group was 3 years.

ID	Year of diagnosis	Group	Time since diagnosis (years) when recruited
210717_005	**2017**	**Drug Naive**	**0**
210717_006	**2017**	**Drug Naive**	**0**
210717_016	**2013**	**Drug Naive**	**4**
210717_019	**2017**	**Drug Naive**	**0**
210717_022	**2016**	**Drug Naive**	**1**
210717_023	**2015**	**Drug Naive**	**2**
210717_025	**2017**	**Drug Naive**	**0**
210717_028	**2013**	**Drug Naive**	**4**
210717_030	**2016**	**Drug Naive**	**1**
210717_033	**2017**	**Drug Naive**	**0**
210717_038	**2015**	**Drug Naive**	**2**
181017_009	**2017**	**Drug Naïve**	**0**
181017_016	**2017**	**Drug Naive**	**0**
181017_019	**2017**	**Drug Naive**	**0**
181017_020	**2005**	**Drug Naive**	**12**
181017_021	**2017**	**Drug Naive**	**0**
181017_023	**2017**	**Drug Naive**	**0**
181017_024	**2016**	**Drug Naive**	**1**
191017_003	**2016**	**Drug Naive**	**1**
191017_004	**2017**	**Drug Naive**	**0**
191017_007	**2015**	**Drug Naive**	**2**
181017_004	**2002**	**Medication**	**15**

181017_005	**2016**	**Medication**	**1**
181017_006	**2014**	**Medication**	**3**
181017_010	**2016**	**Medication**	**1**
181017_012	**2016**	**Medication**	**1**
181017_013	**2013**	**Medication**	**4**
181017_018	**2013**	**Medication**	**4**
191017_006	**2007**	**Medication**	**10**
191017_008	**2015**	**Medication**	**2**
191017_009	**2014**	**Medication**	**3**
210717_009	**2014**	**Medication**	**3**
210717_012	**2016**	**Medication**	**1**
210717_018	**2014**	**Medication**	**3**
210717_029	**2014**	**Medication**	**3**
210717_031	**2015**	**Medication**	**2**
210717_032	**2015**	**Medication**	**2**
210717_034	**2017**	**Medication**	**0**
210717_035	**2014**	**Medication**	**3**
210717_037	**2004**	**Medication**	**13**



No. of Variables	Area Under Curve (AUC) at 95% Confidence Interval	Confidence Interval Range
2	72.0 %	37.8 – 94.2 %
3	72.8 %	47.5 – 91.9 %
4	74.9 %	50.9 – 91.4 %
5	74.3 %	47.5 – 91.9 %
6	76.2 %	53.8 – 92.3 %
7	76.4 %	50.4 – 93.8 %
8	77.2 %	52.4 – 93.8 %
9	77.6 %	47.5 – 93.4 %



No. of Variables	Area Under Curve (AUC) at 95% Confidence Interval	Confidence Interval Range
2	72.4 %	43.9 – 90.5 %
3	76.5 %	51.1 – 90.5 %
4	78.7 %	53.8 – 97.1 %
5	80.3 %	58.6 – 97.6 %
6	82.3 %	57.1 – 97.6%
7	82.8 %	55.8 – 100 %
8	84.8 %	57.7 – 100 %
9	84.3 %	56.4 – 100 %

Figure S1: ROC plots generated using all nine metabolites that were common between the two cohorts (A discovery and B validation) (but not necessarily differential using Student's t-test or expressed in the same direction between cohorts). Each model was built using PLS-DA to rank all variables and top two important variables were selected to start with. Then in each subsequent model additional variables by rank were added to generate ROC curve. Confidence intervals were calculated by Monte Carlo Cross Validation (MCCV) using balanced sub-sampling with multiple repeats.

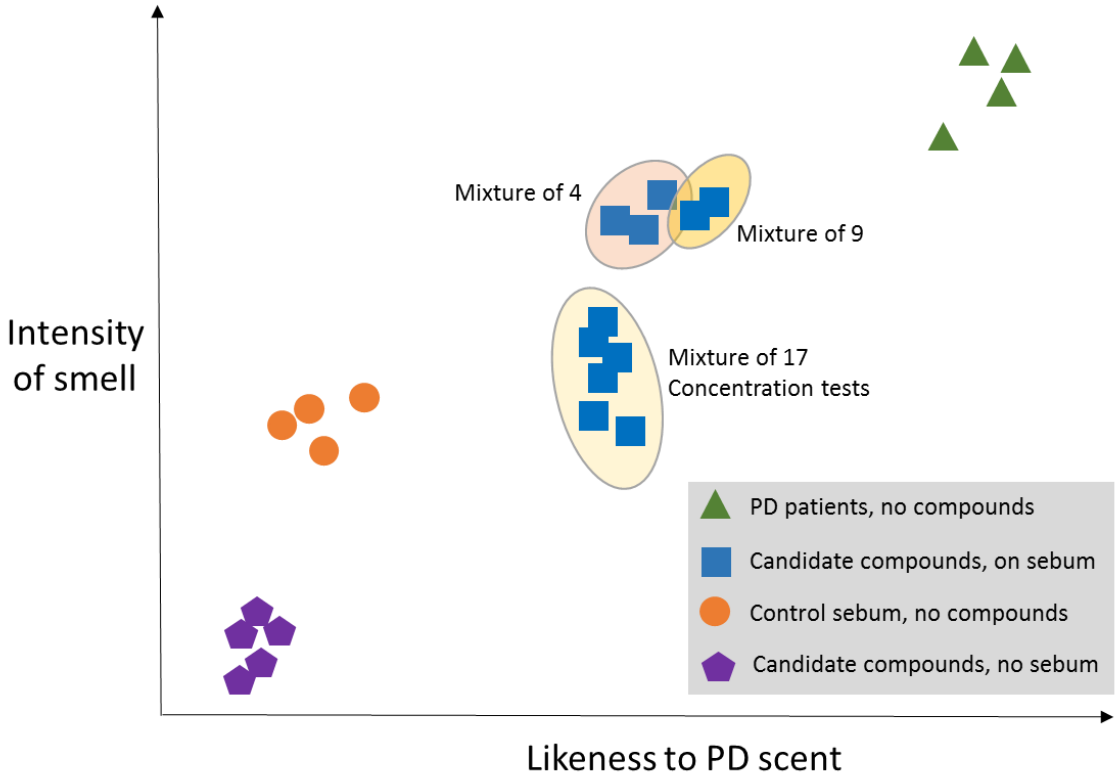


Figure S2: Schematic to show qualitatively the results of a series of blind randomised studies performed by the Super Smeller to classify and score samples based on their similarity to the 'PD smell' and their overall intensity of smell, these scores were defined by both oral and physical denomination by the Super Smeller. All samples were presented on gauze swabs; **purple** pentagons display a cluster of swabs of multiple combinations of candidate compounds spiked onto blank gauze (no sebum), **orange** circles show gauze swabs containing only human control sebum (no compounds), **blue** squares depict three series of compound combinations based on MS analysis spiked on to control sebum and **green** triangles represent clinical gauze samples swabbed from PD patients.